# Accelerating Large Language Model Decoding with Speculative Sampling

*Year: 2023* | *Citations: 627* | *Authors: Charlie Chen, Sebastian Borgeaud, G. Irving*

---

## Abstract

We present speculative sampling, an algorithm for accelerating transformer decoding by enabling the generation of multiple tokens from each transformer call. Our algorithm relies on the observation that the latency of parallel scoring of short continuations, generated by a faster but less powerful draft model, is comparable to that of sampling a single token from the larger target model. This is combined with a novel modified rejection sampling scheme which preserves the distribution of the target model within hardware numerics. We benchmark speculative sampling with Chinchilla, a 70 billion parameter language model, achieving a 2-2.5x decoding speedup in a distributed setup, without compromising the sample quality or making modifications to the model itself.